Evaluating the Performance of the LI3P in Latent Profile Analysis Models

by

Russell Houpt

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved April 2022 by the
Graduate Supervisory Committee:

Kevin J. Grimm, Chair
Daniel McNeish
Michael C. Edwards

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

Latent profile analysis (LPA), a type of finite mixture model, has grown in popularity due to its ability to detect latent classes or unobserved subgroups within a sample. Though numerous methods exist to determine the correct number of classes, past research has repeatedly demonstrated that no one method is consistently the best as each tends to struggle under specific conditions. Recently, the likelihood incremental percentage per parameter (LI3P), a method using a new approach, was proposed and tested which yielded promising initial results. To evaluate this new method more thoroughly, this study simulated 50,000 datasets, manipulating factors such as sample size, class distance, number of items, and number of classes. After evaluating the performance of the LI3P on simulated data, the LI3P is applied to LPA models fit to an empirical dataset to illustrate the method's application. Results indicate the LI3P performs in line with standard class enumeration techniques, and primarily reflects class separation and the number of classes.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Finite mixture models (FMMs) are a series of statistical models often used to identify potentially unmeasured groups within heterogeneous data (McLachlan & Peel, 2000). FMMs have grown in popularity over the past decades due to the integration of FMMs with structural equation models, improvements in computational power, and theories highlighting differential effects. The unmeasured groups in an FMM, called *latent classes* or *latent profiles*, are often assumed to represent homogenous subpopulations from a larger heterogenous population. FMMs are the foundation of two popular classification techniques: latent class analysis (LCA) and latent profile analysis (LPA). LCA was originally implemented to detect latent classes with ordinal variables, whereas LPA was used when dealing with continuous indicators. More recently, these methods have been extended beyond their original uses to accommodate both ordinal and continuous variables, as well as count and nominal variables, and any combination of these variable types (Magidson & Vermunt, 2002). Moreover, FMMs have been combined with a variety of statistical models, such as regression analyses (e.g., Liu, & Lin, 2014), factor models (e.g., Lubke & Muthén, 2005), and growth models (Muthén & Shedden, 1999) to search for unobserved groups with different model parameters.

The process of using FMMs in empirical data involves fitting several models with differing numbers of latent classes and different parameter constraints across classes. For example, a researcher who is exploring empirical data for latent classes with an LPA may fit an LPA with zero covariances within each class for k successive classes, and then fit an LPA with estimated covariances within each class for k successive classes. The fit of

each LPA model is recorded using several different techniques and after all of the models are run, the fit of the models is compared to determine the optimal model. Additionally, the researcher may consider the interpretability of the latent class parameters and the mapping of the latent classes to theoretical expectations when determining which model configuration to use.

The number of subpopulations and their model parameters are unknown with empirical data, which makes determining the correct number of classes a primary challenge. To this end, multiple *class enumeration* techniques have been proposed over the years to aid researchers and have continued to be modified and improved. Recently, a new approach, called the likelihood incremental percentage per parameter (LI3P), was proposed to measure the proportional improvement in model fit (Grimm et al., 2020).

In this thesis, I explore the commonly used fit indices for determining the number of latent classes and review recent research into the accuracy and reliability of these techniques for LCA and LPA models. I then examine new LI3P measure and evaluate its performance relative to the other commonly implemented fit indices using a series of Monte Carlo simulations. I conclude by using the LI3P on empirical data from a recent study that used LPA models to search for latent classes among individuals who no longer believe in God but the standard approaches failed to yield clear results.

Class Enumeration Techniques

Three broad approaches exist for the current primary class enumeration methods. The first, information criteria (IC), refers to a group of techniques that seek to quantify model fit while penalizing overly complex models. The second approach is a group of likelihood ratio tests (LRTs) that involve using a chi-square distribution (or an

approximation) to determine if the model fit significantly improves when a latent class is added. The third approach is a group of resampling techniques, such as *k*-fold cross validation. Here, I review the theory and performance of the first two methods as they are the most commonly used and studied.

Information Criteria. ICs attempt to quantify the fit of the model while penalizing complexity. To achieve this, the -2 log likelihood ($-2LL$) is used to quantify model fit with a penalty imposed for each estimated parameter. Likelihood is a function used to determine the probability of the observed data given a specified model. For LCA and LPA models, the likelihood function refers to the calculated probability that the observed sample data is drawn from a population described by the estimated model with its specific number of classes and parameter estimates, with values ranging between zero and one. The natural log of the likelihood function is taken to transform the likelihood values to larger values to improve computational efficiency. Finally, this log likelihood is multiplied by -2, transforming the values into positive numbers. Altogether, -2 log likelihood can be calculated as

$$-2LL = -2 \cdot \sum_{i=1}^{N} \left(-\frac{K_i}{2} \ln(2\pi) - \frac{1}{2} \log|\mathbf{\Sigma}_i| - \frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_\mathbf{i})' \mathbf{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right) \quad (1)$$

where i identifies the case, K_i is the number of variables for each case, $|\mathbf{\Sigma}_i|$ is the determinant of the model-implied covariance matrix for case i, $\mathbf{y}_i$ is the vector containing the scores for case i, and $\boldsymbol{\mu}_i$ is the model implied mean vector for case i. Thus, $-2LL$ values close to zero represent a good fitting model, whereas higher values indicate worse fitting models and greater discrepancy between the observed empirical data and the model-implied covariance matrix and mean vector, holding the data constant (Grimm et

al., 2020).

ICs take the $-2LL$ and apply a penalty in an effort to avoid overly-complex models. There are two common IC groups: the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978).

The AIC is calculated as

$$\text{AIC} = -2LL + 2p \tag{2}$$

where -2*LL* is the -2 log likelihood equation and *p* is the number of estimated parameters in the model. The AIC has a constant penalty for each estimated parameter, which tends to under penalize models fit to data with moderate or larger sample sizes. Additionally, the AIC is inconsistent, meaning that the AIC will not indicate the correct model as the number of parameters increases to infinity (Woodroofe, 1982). However, despite being inconsistent, it has still been found to be useful in certain situations (e.g., Bozdogan, 1987; Gonzalo & Pitarakis, 2002).

Recognizing the issue with the AIC in larger samples, Hurvich and Tsai (1989) proposed another form of the AIC, known as the HT-AIC or AICc, which adds an additional term to the AIC formula in order to adjust for sample size. The AICc can be calculated as

$$\text{AICc} = \text{AIC} + \frac{2p^2 + 2p}{N - p - 1} = -2LL + 2p + \frac{2p^2 + 2p}{N - p - 1} \tag{3}$$

where *p* is the number of parameters and *N* is the sample size. Additional versions of the AIC have been proposed, such as the consistent AIC (CAIC; Bozdogan, 1987) and the AIC3 (Andrews & Currim, 2003); however, these forms are not commonly implemented or studied.

The second popular IC group is the BIC, which is calculated as

$$\text{BIC} = -2LL + \ln(N) \cdot p \quad (4)$$

where -2*LL* is the -2 log likelihood equation, *N* is the sample size, and *p* is the number of parameters in the model. Like the AIC, the BIC applies a penalty to the -2*LL* based on the number of estimated parameters, but the penalty term used here is not constant. Instead, the penalty parameter is scaled by the natural log of the sample size, resulting in a larger penalty than the AIC when $N > 7$. Thus, the BIC tends to select models with fewer parameters compared to the AIC. It is due to this non-linear penalty term that the BIC is considered to be consistent, unlike the AIC.

As with the AIC, several modifications of the BIC have been proposed. Sclove (1987) proposed a sample size adjusted form (SaBIC) replacing *N* in the BIC formula with an altered form, calculated as

$$\text{SaBIC} = -2LL + ln(\frac{N+2}{24}) * p \quad (5)$$

which lowers the penalty term in the BIC, making it prefer relatively more parameterized models. The saBIC has often been shown to do well compared to the BIC, though it needs 50 or more cases per class to perform well (Yang, 2006). This is perhaps the most popular derivation of the BIC that is used, though others exist, such as the DIC (Draper, 1995) or the ICL-BIC (Biernacki et al., 2000), though these are used less commonly and have rarely been investigated in FMM simulations.

Likelihood Ratio Tests. A second class of enumeration technique approach are the LRTs, which attempt to use a chi-square distribution to determine improvements in model fit for nested models. In FMMs, LRTs are implemented by comparing a model

with k classes to a model with $k-1$ classes given the same set of parameter constraints, such that the latter model is nested under the former. With this assumption, a chi-square distribution could then be used to determine if the k class model fits the data significantly better than the $k-1$ model. However, the change in $-2LL$ does not follow a chi-square distribution because the parameter constraints applied to the k class model to create the $k-1$ class model are on the boundary of the parameter space (McLachlan & Peel, 2000).

An alternate LRT method was proposed by Lo, Mendell, and Rubin (LMR-LRT or VLMR-LRT, 2001), who developed an approximation to the chi-square distribution for the difference between the $-2LL$ values. This was based on the work done by Vuong (1989) and allowed researchers to statistically compare the k and $k-1$ class models and produce a *p*-value to check for improved model fit. Lo, Mendell, and Rubin (2001) also proposed an ad hoc adjustment to this test (aLMR-LRT), which scales the LMR-LRT as a function of the parameters so it converges to the correct Type I error rate faster, though in practice both tests yield nearly identical results (Tofighi & Enders, 2008; Peugh & Fan, 2013). In 2003, the LMR-LRT was found to have a mathematical flaw in the underlying proof by Jefferies (2003). Despite this flaw, several simulation studies have found the LMR-LRT to be useful in determining the correct number of latent classes (Lo et al., 2001; Olivera-Aguilar & Rikoon, 2018).

A second LRT approach, the parametric bootstrap likelihood ratio test (BLRT), was developed by McLachlan and Peel (2000). This test also compares the fit of the *k* and *k*-1classes, but does so by recording the difference in -2*LL* and simulating data repeatedly based on the parameter estimates of the *k*-1 class model. The *k* and *k*-1 models are then fit to these simulated datasets and the difference in -2*LL* values are used to create a sampling

distribution. The -2*LL* difference obtained from the empirical dataset is then compared to this distribution of *-2LL* values yielding a *p*-value to determine if the model fit is significantly better. This method has grown in popularity, but it has not been thoroughly studied because the method takes a significant amount of computational time compared to other methods (Nylund et al., 2007).

Review of Class Enumeration Performance in Past Studies

Previous research has explored the performance of these class enumeration techniques using simulation methods. The research reviewed here includes both LCA and LPA models, as they are conceptually alike and the findings tend to be similar, with a focus on simulation conditions, notable results, and trends in the performance of the indices.

Latent Class Analysis Studies. In 2006, Yang conducted an LCA simulation study which focused on the performance of six ICs while manipulating three factors: sample size, the number of binary indicator variables, and the number of latent classes. Yang found that sample size affected IC performance, with smaller samples decreasing accuracy. Yang also found that higher numbers of latent classes lowered IC accuracy. In general, the AIC was found to be unreliable and overestimated the number of classes, though it did perform well with small sample sizes and many classes. The CAIC performed slightly better, but was even more unreliable. The BIC generally underestimated the number of classes but performed well with large sample sizes. The saBIC performed the best in this study, with the highest overall accuracy of the ICs examined.

A follow-up study by Yang and Yang (2007) found similar results. In this study,

sample size, the number of classes, and three different class configurations were manipulated. As before, smaller sample sizes and more classes were associated with decreased IC accuracy. The AIC again performed well with small samples and many classes, whereas the BIC performed poorly with a large number of classes and low sample sizes. Notably, the modified ICs that used a sample size adjustment (such as the saBIC) performed better than the original form. These results were mirrored by Zhang et al. (2014), whose research also focused on the use of LCA with binary variables.

Swanson et al. (2012) also explored class enumeration in LCA models, but manipulated additional conditions including the number of indicators per class, conditional probabilities, missing data, local dependence, and sample composition. Using these conditions, they found that most ICs improved with increasing sample size. They also found that the AIC performed poorly for almost every condition, again overestimating the true number of classes. As before, the BIC often underestimated the number of classes, whereas the saBIC again performed the best overall.

Nylund et al. (2007) examined LCA models using both binary and continuous indictors, which overlaps with the typical application of LPA models. This simulation manipulated several factors, including the number of items, sample size, and the number of latent classes. Again, the AIC struggled to find the correct number of classes, often overestimating the correct number. The BIC achieved 100% accuracy in some conditions, but only 8% in others, while the saBIC was more stable but not as accurate, often overestimating the number of classes. The authors thus recommend the BIC as the best performing IC with the caveat that it struggles when $N < 500$. This study also examined LRTs, finding that the LMR-LRT performed well but not nearly as well as the BLRT,

which the authors concluded was the best performing class enumeration method, though they warned it increased computation time by 5 to 35 times, so they ultimately recommended using it to confirm the number of classes after using the BIC and LMR-LRT.

Latent Profile Analysis Studies. Similar conclusions were found in simulation research on model comparison statistics in LPA research. Tein et al. (2013) performed a study examining LPA models and class enumeration techniques where they manipulated sample size, the number of identifiers, inter-class distance, and the number of classes. They found that sample size had a smaller impact than the number of indicators, finding that more indicators increased power for certain ICs and LRTs. They also found that inter-class distance was more influential than the number of classes, such that greater distances led to greater power in the techniques. The AIC had low power for most conditions and often overestimated the number of classes, whereas the BIC and saBIC generally performed well, though the BIC did favor models with fewer classes. The LMR-LRT and aLMR-LRT performed well, but only when given a medium or large sample size. Notably, they found that there was not reliable power to detect the number of classes no matter the method used when *d* was .2 or .5.

Peugh and Fan (2013) manipulated several factors in their study, including sample size, Mahalanobis distance (latent class separation), the number of indicator variables, and the number of latent classes. They discovered many of the ICs and LRTs struggled in the 1-class condition, with the IC derivatives performing best and a notable increase in performance after overriding the default local independence and homogeneity assumptions. In the 3-class case, the saBIC performed best, though it still struggled

significantly and many of the ICs and LRTs achieved 0% accuracy. These results fall in line with Yang (2006) and Nylund et al. (2007) in terms of relative performance, though the accuracies were significantly lower. The authors suggested that the poor performance was potentially due to the small sample size, citing Paxton et al. (2001) who suggests that a sample of less than 500 should be considered small, as well as model misspecification due to unmet local independence assumptions.

Morgan et al. (2016) manipulated the sample size, the number of indicators, the shape of the distribution of the indicators, and profile proportions for their study. This resulted in 67,500 datasets, which were analyzed twice – once using the original data and a second time where the data was transformed to be normally distributed (using van der Waerden quantile normal scores). The AIC performed poorly overall, with the distribution of indicators being a major factor in its performance. The BIC and the saBIC performed the best among the ICs, particularly with normally distributed indicators and when there were more than ten; the saBIC only correctly identified the number of latent classes once in the non-normal data. The LMR-LRT was less affected by non-normality in the indicators, often performing as well as or better than the ICs. However, once the data was transformed to be normally distributed, the BIC and saBIC accuracy increased and surpassed that of the LMR-LRT.

Olivera-Aguilar and Rikoon's (2018) simulation study examined LPAs and manipulated sample size, the proportion of the sample in each class, latent profile size, the magnitude of invariance violations, and the number of violating indicators. They found that the LRT outperformed the ICs and that the AIC outperformed the BIC and saBIC. These results are unique compared with the other literature, but they do fall in line

with Finch (2015) who also examined noninvariance. Wang et al. (2021) performed a similar study, but found results that were more in line with the rest of the literature, potentially, as they suggest, due to differences in how measurement noninvariance was defined in their respective studies.

In summary, several distinct trends emerge in class enumeration research. The AIC appears to perform well in small samples and less-separated classes, but often overestimates the number of classes in larger samples. The BIC tends to perform well in larger samples and better-defined classes, but often underestimates the number of classes in smaller samples. The saBIC generally performs better than the BIC but seems to struggles under certain conditions, such as non-normality. Finally, despite the LMR-LRT's mathematical flaw, it still performs well – sometimes better than any of the ICs – though it too struggles in small samples.

More generally, it is clear that sample size is often influential in the performance of the class enumeration methods. Several researchers have suggested that samples below 500 may be too small to draw reliable results (e.g., Paxton, 2001; Nylund et al., 2007). Typically, the larger the sample, the more accurate the results (Swanson et al., 2012), though some studies found that certain ICs struggle in these circumstances (e.g., Yang & Yang, 2007). For ICs, this is certainly important as the -2*LL*, the key measure of model fit, changes linearly with sample size (Grimm et al., 2021). Likewise, LRTs are also affected by sample size, often struggling to make the correct determinations with small samples, whereas large samples can make almost meaningless small differences appear important. The importance of adequate sample size cannot be understated, and the issue is not always easily resolved, especially when working with hard to obtain samples.

There are several other important factors, including the number of classes, distance between classes, and the number of indicators. Several studies demonstrated that increasing the number of classes decreases the accuracy of the techniques, with the exception of the BIC. The distance between classes can be extremely challenging for the LPA, so much so that Tein et al. (2013) posited that distances smaller than $d = 0.5$ may be too small to reliably detect. Finally, the number of indicators has repeatedly been demonstrated to be influential, with more indicators making it easier for the techniques to correctly recover the number of simulated classes.

Given these findings, it is not surprising that researchers may find themselves in complicated situations when working with empirical data. Class enumeration methods, such as the BIC, the saBIC, or the LMR-LRT, may seem to be more reliable, but there exist numerous conditions where they may perform poorly and mislead researchers. This has led researchers to suggest using multiple methods together and examining the results of each to get a comprehensive understanding of what is happening (e.g., Nylund et al., 2007; Ram & Grimm, 2012). Even then, researchers often find themselves in situations where some techniques used indicate one model configuration while the rest indicate another (e.g., McLaughlin et al., 2020). Here, a researcher would then need to consider the theoretical and interpretive significance of choosing one number of classes over the other – a precarious situation, especially if the analysis is exploratory or if the results do not align with the initial hypotheses.

The LI3P. Grimm et al. (2021) recently proposed and briefly explored a new method, the likelihood incremental percentage per parameter (LI3P), which takes a different approach to comparing models. The LI3P method, a modified form of likelihood incremental

percentage (LIP) method proposed by McArdle et al. (2002), is not a new IC, LRT, or a resampling method, but rather a sort of effect size for determining the amount of improvement in relative model fit.

The LIP is calculated as

$$\text{LIP} = 100 \cdot \left(1 - \frac{-2LL_{\,more\ parameterized\ model}}{-2LL_{\,less\ parameterized\ model}}\right) \tag{6}$$

which is similar in structure to a pseudo r-square effect size in logistic regression. The LIP contextualizes the change in the model fit by dividing the *k* class model's -2*LL* by the *k*-1 class model's -2*LL*, returning a percentage value for model fit improvement. Grimm et al. (2021) proposed an altered form, the LI3P, after their initial analysis, defined as

$$\text{LIP} = 100 \cdot \left(1 - \frac{-2LL_{\,more\ parameterized\ model}}{-2LL_{\,less\ parameterized\ model}}\right)/p \tag{7}$$

where *p* is the number of additional parameters from the previous and current model. This scales the LIP value so that it can be compared to other models after removing the effects of additional parameters. Like other effect size measures, this method would, in theory, be less affected by sample size and help provide much-needed information in ambiguous situations.

Grimm et al. (2021) initially tested this measure in a series of LPA models, replicating the conditions used by Nylund et al. (2007). The LI3P values seemed to reliably indicate the correct number of classes, with larger values being reported as the models being fit approached the true number of classes, followed by a large drop in LI3P values for models with unneeded additional latent classes. The authors cautioned against using the LI3P in samples when $N < 500$, stating that the LI3P values were larger than

expected, potentially indicating bias. They also proposed some effect size cutoffs, suggesting a score of 0.1 was small, 0.1 to 0.3 was medium, and 0.3 or larger was large.

CHAPTER 2

METHODS

To test the performance of the LI3P in LPA models more thoroughly, the LI3P was calculated for a series of LPA simulations using various conditions that include challenging configurations of sample size and latent class separation. According to a review by Morgan et al. (2016), the most common number of classes chosen by empirical studies is three, so data was simulated with one, two, or three classes. Tien et al. (2013) found that most techniques struggle to perform when inter-class distance, measured as Cohen's *d*, is less than .5, so mean differences were chosen between .25 and 1.0, randomly selecting values for each simulation rather than using a set of discrete values. Likewise, sample sizes were randomly chosen, such that the minimum total number of samples across classes was 300 and the maximum total number of samples was 1,500. Values were chosen this way in an effort to achieving a more continuous perspective on how the LI3P performs while creating a unique data set that lends itself to novel analyses for this area of simulations.

Fifty thousand datasets consisting of 15 variables were created in R, drawing from the conditions described above, with 1-3 classes randomly chosen as the true number of classes and a total sample size between 300 and 1500. The first class was always centered at 0; with subsequent class means randomly chosen to be 0.25 to 1.0 units away from previous class, such that $\mu_1 < \mu_2 < \mu_3$. The latent profile analysis models were then fit in M*plus* (version 7.4; Muthén & Muthén, 1998-2017). For each 15-variable dataset, 12 models were fit: first, LPA models with one through four classes were fit to the 15 variables (one more than the maximum number of classes simulated), then the process

was repeated on a subset of 10 variables, then five. This process yields 150,000 datasets (50,000 by 3 variable configurations), each with 4 LPA models fit to them. Popular class enumeration techniques were used for performance comparison, including the AIC, saBIC, and the aLMR-LRT. The results were imported back into R, where the LIP and LI3P were calculated.

CHAPTER 3

RESULTS

The accuracy of class enumeration techniques, separated by the number of classes and the number of indicators, the categorical manipulations, are reported in Table 1, Table 2, and Table 3, where Table 1 displays the results for when there was one class in the population, Table 2 contains the results for then there were two classes in the population, and so on. Each row is identified by the number of indicators analyzed, whereas the columns identify the index used followed by the number of classes. For example, row "5" and column "SaBIC 1" in Table 1 reports the percentage of times the SaBIC chose one class as the solution across simulations, in this case 98. The correct number of classes for each table bolded.

Percentages were calculated after removing models that produced errors or warnings, so the values reported in the tables reflect models only for which a researcher would presumably accept the values as valid. In some cases, a simulation's data was completely removed because every model failed to converge or failed to converge to a proper solution (e.g., negative variance) as indicated by M*plus* warnings. Models may fail to converge for a variety of reasons, including attempting to fit models to data that were greatly misaligned (e.g., fitting a four-class model to a one-class data set), or when the distances between classes were too small to properly detect the number of classes. Of the 600,000 LPA models fit, 46.55% were flagged and were not used in the analyses. Of these flagged models, only 7% were models fitting the correct number of classes; the remaining 93% occurred almost entirely in models that to fit too many classes.

To compare the performance of the LI3P to the other techniques, the LI3P values

Table 1: Model selection for the one-class condition

Number	AIC				SaBIC				aLMR-LRT				LI3P			
of Vars	**1**	2	3	4	**1**	2	3	4	**1**	2	3	4	**1**	2	3	4
5	**63.7**	28.4	6	1.9	**98**	1.9	0.1	0	**98.9**	1	0.1	0	**67.1**	29.4 (0.021)	3.4 (0.02)	0.1 (0.018)
10	**71.8**	23.6	3.3	1.3	**99.5**	0.4	0	0	**100**	0	0	0	**93.1**	6.5 (0.011)	0.3 (0.01)	0 (0.011)
15	**79.3**	17.8	2.1	0.7	**99.9**	0.1	0	0	**100**	0	0	0	**98.9**	1.1 (0.007)	0 (0.007)	0 (0.006)

Each cell reports the % of simulations under the variable condition that the method chose the corresponding column class solution. Bolded values indicate the simulated condition. For the LI3P, the number in parentheses is the average LI3P value for models chosen with that solution.

Table 2: Model selection for the two-class condition

Number	AIC				SaBIC				aLMR-LRT				LI3P			
of Vars	1	**2**	3	4	1	**2**	3	4	1	**2**	3	4	1	**2**	3	4
5	13.1	**60.5**	22	4.4	34.7	**64.5**	0.8	0	41.3	**57.9**	0.7	0.1	13.5	**63.4 (0.074)**	21.4 (0.019)	1.7 (0.018)
10	7.1	**72**	17.4	3.6	23	**76.9**	0.1	0	27.8	**72.1**	0.1	0	21.1	**76.7 (0.062)**	2.1 (0.01)	0.1 (0.009)
15	4.4	**73.8**	19.2	2.7	17.1	**82.8**	0	0	20.6	**79.3**	0.1	0	24.9	**75 (0.052)**	0.2 (0.006)	0 (0.006)

Each cell reports the % of simulations under the variable condition that the method chose the corresponding column class solution. Bolded values indicate the simulated condition. For the LI3P, the number in parentheses is the average LI3P value for models chosen with that solution.

Table 3: Model selection for the three-class condition

Number	AIC				SaBIC				aLMR-LRT				LI3P			
of Vars	1	2	**3**	4	1	2	**3**	4	1	2	**3**	4	1	2	**3**	4
5	0.4	51.7	**39.6**	8.2	2.4	89.1	**8.4**	0.1	4.8	88.8	**6.2**	0.2	0.5	53.5 (0.074)	**40.5 (0.021)**	5.6 (0.017)
10	0.1	39.7	**49.9**	10.4	0.4	71.6	**27.9**	0	1.1	74.9	**23.9**	0	0.4	70.8 (0.062)	**28.4 (0.018)**	0.4 (0.009)
15	0	26.4	**64.9**	8.7	0.1	58.1	**41.8**	0	0.4	61.4	**38.1**	0	0.5	69.8 (0.052)	**29.8 (0.016)**	0 (0.006)

Each cell reports the % of simulations under the variable condition that the method chose the corresponding column class solution. Bolded values indicate the simulated condition. For the LI3P, the number in parentheses is the average LI3P value for models chosen with that solution.

for the one-class, 15-variable condition were examined and the 95th percentile value, 0.014, was chosen as a cutoff score. This was done because the LI3P, like other effect size measures, is not directly made for decision making as there are no explicit decision criterion. Using the 95% value was done to approximate the 95% accuracy that most statistical tests strive to achieve and was tuned to the one-class case as it was both computationally easy and conceptually functions as the null.

The aLMR-LRT performed the best in the one-class case, whereas the AIC was consistently the worst. In the two-class condition, the SaBIC consistently performed best,

though in this condition the methods difference between the methods was smaller. There was no clear worst comparison approach in the two-class condition; the aLMR-LRT performed the worst at the five-indicator condition, whereas the AIC performed the worst at the 10 and 15-indicator conditions. Table 3 shows that the aLMR was the worst performing for the five and 10 variable conditions, but the LI3P was the worst performing in the 15-variable condition.

Specifically looking at the LI3P's performance, the method performed slightly better than the AIC measure in the one-class case five variable, but performed nearly as well as the SaBIC measure in the 15-variable case. In the two-class case, the LI3P was the second-best performing approach used here, just behind the SaBIC, though the 15-variable condition it had fallen to third-best. Notably, it performed better in the 10-variable condition than the 15-variable condition, possibly indicating that the abundance of information might confound the technique. Supporting this idea, in the three-class condition, the LI3P's performance was best in the five-variable condition, with both the LI3P and the AIC vastly outperforming the SaBIC and the aLMR-LRT. In the 15-variable condition, the AIC was the best-performing technique, and the LI3P was now the worst.

The Adjusted Rand Index (ARI), which evaluates accuracy after adjusting for chance (Hubert & Arabie, 1985), was used to determine the overall accuracy of the class enumeration techniques across class conditions, broken down by the number of variables. These results are displayed in Table 4. Examining Table 4, it becomes alarmingly clear that none of the class enumeration techniques were truly reliable. Across the variable condition, the best accuracy was less than 50% and the worst was under 30%; by

Table 4: Adjusted Rand Index values

Number of Indicators	AIC	SaBIC	aLMR-LRT	LI3P
5	16.82	39.81	38.16	18.68
10	27.72	44.81	42.93	38.85
15	40.24	50.29	47.56	43.48
Overall	28.26	44.97	42.88	33.67

The ARI values reveal the accuracies across class conditions, adjusting for chance. The values in the "Overall" row show the overall accuracy for the measures, adjusting for chance.

examining the effects of the number of indicators, this can be improved to a best accuracy of just over 50%, but the worst accuracy sits at less than 20%. The variability was also worth considering; the range in AIC accuracy changed was 23.4%, 10.5% for the SaBIC, 9.4% for the aLMR-LRT, and 24.8% for the LI3P. Based on these results, the SaBIC and the aLMR-LRT seemed to perform the best, both having relatively high overall accuracy with relatively low variability in performance. The AIC performed poorly, with the lowest overall accuracy and the second greatest range. Interestingly, the LI3P had the largest range in performance and the third-best accuracy. Notably, the order of approaches by accuracy was the same across the variable conditions with the SaBIC performing best and the AIC performing the worst.

Finally, the LI3P was directly compared with each of the three class enumeration techniques using a series of crosstables, which compared the ultimate "decision" of each technique in Tables 5, 6, and 7. These tables compared the overall performance of the two techniques (e.g., the LI3P and the SaAIC), broken down by the number of indicators. Examining these tables highlights that the SaBIC tended to favor more classes compared to the LI3P, but they agreed fairly often, roughly 92.65% of the time in the 15-variable case, though that agreement fell as the number of indicators decreased. The LI3P and the SaBIC only agreed 62.75% of the time in the five-variable condition. The AIC, by

Table 5: Crosstable comparing the SaBIC and the LI3P

		5 Variables SaBIC				10 Variables SaBIC				15 Variables SaBIC			
		1	2	3	4	1	2	3	4	1	2	3	4
LI3P	1	13535	0	15	0	18918	238	2	0	19356	1412	0	0
	2	7828	16541	0	3	1559	23723	365	2	216	22061	2001	0
	3	1162	8377	1299	0	79	799	4219	0	8	34	4910	0
	4	67	929	229	15	8	38	46	4	1	0	1	0
		Agreement: 63%				Agreement: 94%				Agreement: 93%			

Table 6: Crosstable comparing the AIC and the LI3P

		5 Variables AIC				10 Variables AIC				15 Variables AIC			
		1	2	3	4	1	2	3	4	1	2	3	4
LI3P	1	11314	1167	809	260	13017	4896	944	301	13977	5551	1033	207
	2	1498	20777	1136	961	174	17637	6482	1356	11	14170	8850	1247
	3	96	1432	9165	145	3	31	4272	791	0	0	4409	543
	4	4	51	145	1040	0	2	6	88	0	0	0	2
		Agreement: 85%				Agreement: 70%				Agreement: 65%			

Table 7: Crosstable comparing the aLMR-LRT and the LI3P

		5 Variables aLMR-LRT				10 Variables aLMR-LRT				15 Variables aLMR-LRT			
		1	2	3	4	1	2	3	4	1	2	3	4
LI3P	1	13516	11	18	5	18959	192	6	1	19735	1028	5	0
	2	8890	15457	5	20	2436	22874	337	2	478	22234	1564	2
	3	1731	8109	995	3	138	1359	3596	4	22	180	4747	3
	4	102	970	141	27	9	53	31	3	1	1	0	0
		Agreement: 60%				Agreement: 91%				Agreement: 93%			

comparison, agreed with the LI3P 65.11% of the time in the 15-variable case, with the AIC often choosing more classes than the LI3P. In the five-variable condition, that agreement was 82.51%, with the AIC choosing more classes and the LI3P choosing fewer. Finally, the aLMR-LRT and the LI3P were in agreement 93.43% of the time for the 15-variable condition, with the aLMR-LRT occasionally choosing more classes than the LI3P. In the 5-variable condition, the agreement dropped to 59.94%, with the LI3P often choosing fewer classes relative to the aLMR-LRT.

To determine the effects of the continuous manipulations, the sample size and class distances, a series of 12 scatterplots (Figures 1-12) were created. These plots depict the mean difference in the two-class condition on the vertical axis and the total sample size on the horizontal axis. Blue dots represent a simulation where the method correctly

identified the number of classes, and red a simulation where the method was incorrect.

These scatterplots help visually determine important factors for a given technique. For example, the SaBIC seems to be heavily influenced by class distance, but only minorly by sample size, as evident by the almost flat slope in the line dividing the red and blue along the bottom, and the few red dots along the left axis where sample size is at its smallest. Furthermore, the effect of the number of indicators was also important, as the graph for the 15-variable condition indicates more correct class enumerations compared

Figure 1
SaBIC decision scatterplot - 5 variables

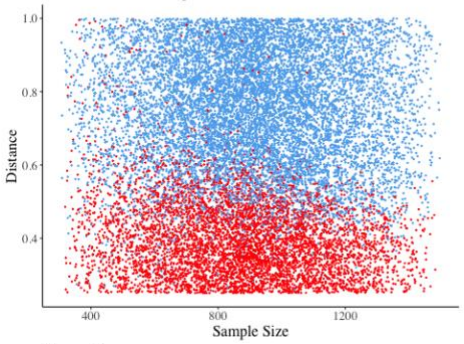


Figure 2
SaBIC decision scatterplot - 10 variables

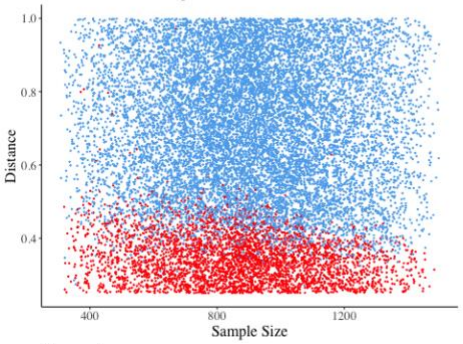


Figure 3
SaBIC decision scatterplot - 15 variables

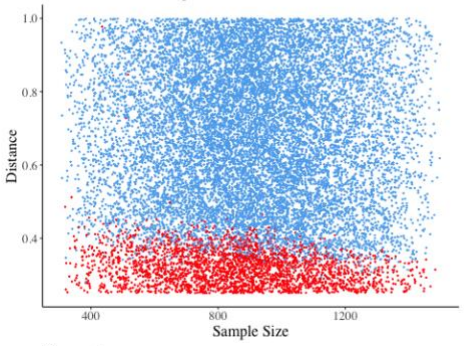


Figure 4
AIC decision scatterplot - 5 variables

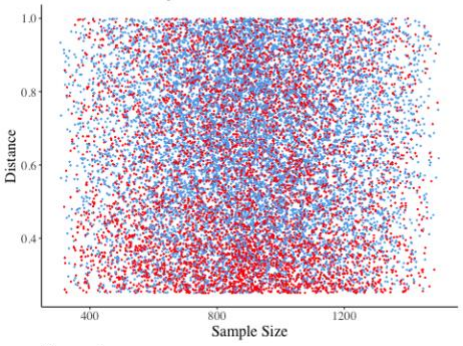


Figure 5
AIC decision scatterplot - 10 variables

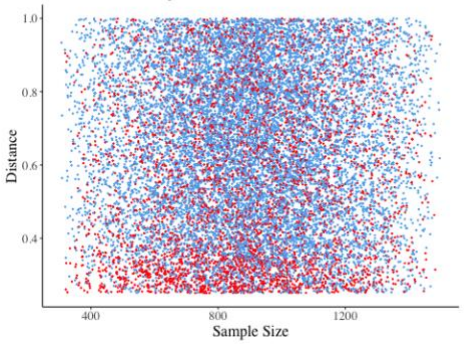


Figure 6
AIC decision scatterplot - 15 variables

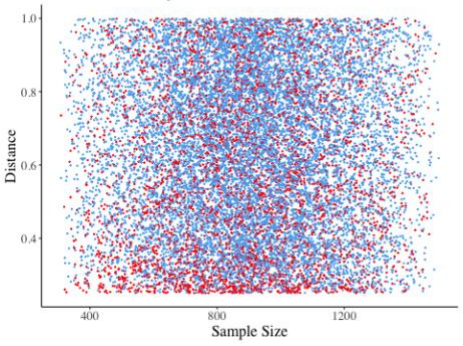


Figure 7
aLMR-LRT decision scatterplot - 5 variables

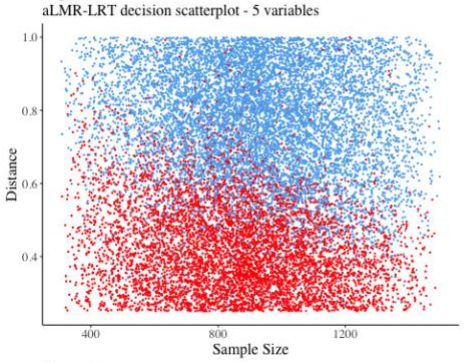


Figure 8
aLMR-LRT decision scatterplot - 10 variables

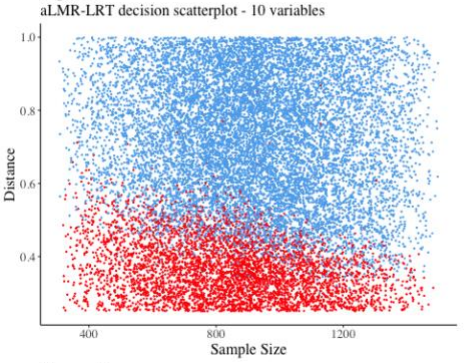


Figure 9
aLMR-LRT decision scatterplot - 15 variables

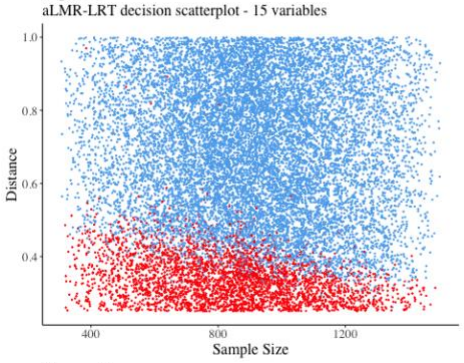


Figure 10
LI3P decision scatterplot - 5 variables

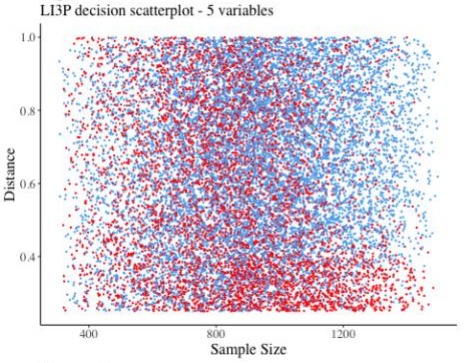


Figure 11
LI3P decision scatterplot - 10 variables

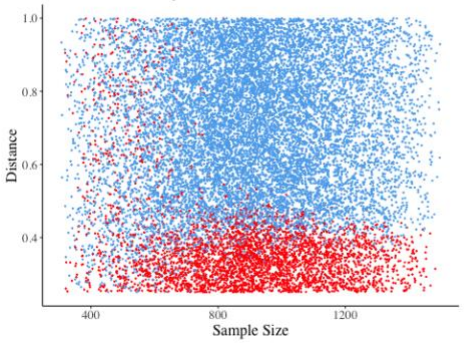
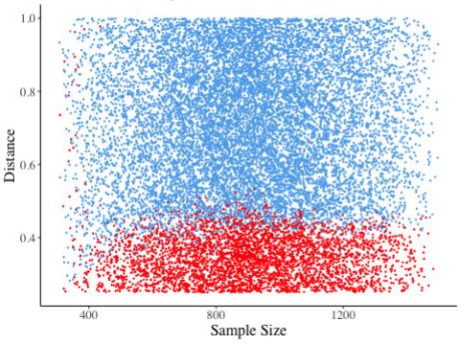


Figure 12
LI3P decision scatterplot - 15 variables



to the five-variable case. By comparison, the AIC does not seem to have any clear relationship between the number of variables and the outcomes as the scatterplot seems almost random. The LI3P, specifically, seems very influenced by the sample size and number of indicators, with the more indicators strengthening the effects of sample size. To further examine the suggested relationships shown in the scatterplots, regression models were fit with the LI3P measure as the outcome and manipulated factors as the predictors. The regression analyses examined the two and three class populations with the

LI3P values predicted from the categorical and continuous conditions manipulated. The estimates of the regression equations are presented in Table 8 and Table 9.

Examining the point estimates in the tables, increasing sample size (scaled to be change per 100) had a small but clear effect. For example, adding 100 observations to the first population in the three-class case would reduce the LI3P score by .002, holding the other conditions constant. Depending on the condition, this raised or lowered the LI3P score. The effect of distance was positive, so an increased distance between classes resulted in a relatively larger change in LI3P values. The effect of the number of variables was rarely important, with the largest effect being class distance.

Examining the squared semi-partial correlations, only the distance between

Table 8: Regression parameters for LI3P scores for the two-class condition

	Dependent Variable:		
Independent Variable:	LI3P score with 5 indicators	LI3P score with 10 indicators	LI3P score with 15 indicators
Intercept	-0.072** (0.001)	-0.079** (0.001)	-0.067** (0.0005)
Sample size of class one	-0.001** (0.0001) 0.0004	0.0002** (0.0001) 0.0001	0.0003** (0.0001) 0.0002
Sample size of class two	-0.001** (0.0001) 0.0007	0.0001 (0.0001) 0.0000	0.0002** (0.0001) 0.0001
Distance between class one and two	0.233** (0.001) 0.7941§	0.216** (0.001) 0.8756§	0.183** (0.0005) 0.9020§
Observations	15,014	15,615	16,020
R^2	0.795	0.876	0.902
Adjusted R^2	0.795	0.876	0.902
Residual Std. Error	0.024 (df = 15010)	0.017 (df = 15611)	0.013 (df = 16016)
F Statistic	F(3, 15010) = 19450.950**	F(3,15611) = 36666.170**	F(3,16016) = 49176.440**

Note: Point estimates are reported with *p* value indicators, followed by standard errors. Squared semi-partial correlation values with their associated *p* value indicators are reported thirdly for predictors.

p<0.01 is indicated with **
Squared semi-partial effect sizes: small (†), medium (‡), large (§)

Table 9: Regression parameters for LI3P scores for the three-class condition

	Dependent Variable:		
Independent Variable:	LI3P score with 5 indicators	LI3P score with 10 indicators	LI3P score with 15 indicators
Intercept	0.022**	-0.020**	-0.033**
	(0.0004)	(0.001)	(0.0005)
Sample size of class one	-0.002**	-0.001**	-0.001**
	(0.0001)	(0.0001)	(0.0001)
	0.0901†	0.0185	0.0090
Sample size of class two	-0.001**	0.001**	0.001**
	(0.0001)	(0.0001)	(0.0001)
	0.0210†	0.0078	0.0144
Sample size of class three	-0.002**	-0.002**	-0.001**
	(0.0001)	(0.0001)	(0.0001)
	0.0956†	0.0225†	0.0078
Distance between class one and two	0.012**	0.032**	0.039**
	(0.0003)	(0.0004)	(0.0003)
	0.0876†	0.2622§	0.3298§
Distance between class two and three	0.013**	0.032**	0.038**
	(0.0003)	(0.0004)	(0.0003)
	0.0894†	0.2671§	0.3289§
Observations	9,365	10,175	12,449
R^2	0.407	0.591	0.667
Adjusted R^2	0.407	0.591	0.667
Residual Std. Error	0.007 (df = 9359)	0.008 (df = 10169)	0.008 (df = 12443)
F Statistic	F(5, 9359) = 1,285.382**	F(5,10169) = 2,937.660**	F(5,12443) = 4,988.680**

Note: Point estimates are reported with *p* value indicators, followed by standard errors, and squared semi-partial correlation values with their associated *p* value indicators.

p<0.01 is indicated with **
Squared semi-partial effect sizes: small (†), medium (§), large (§)

classes had a sizeable effect for the two-class condition, each time being identified as large. In the three-class case, all of the sample sizes and distances had a small effect size in the five-variable condition. The 10-variable condition had a small effect for the third sample size and large effects for the distance. Finally, the 15-variable condition had large effects only for the class distances. Overall, all six of these models were significant, but the R^2 ranged from 0.41 to 0.90, indicating that there is more unexplained variance in the

three-class condition than the two-class condition.

Finally, the `rpart` package (version 4.1.16; Therneaux & Atkinson, 2019) was used to implement classification and regression trees (CART) to determine the effects of the simulation conditions on the LI3P scores, using the Gini index as the splitting criterion and 10-fold cross-validation. The trees were incredibly large, even after pruning and cross-validating, the smallest having 60 nodes and the largest tree having 130, which wasn't entirely unexpected given the criterion is a continuous, not categorical, outcome, and the sample size is extremely large. In general, the number of groups and the distance between groups tended to be selected to initially partition the data, with factors relating to sample size and the ratios between sample sizes appearing lower in the trees. The variable importance metric was extracted from the models and is reported in Table 10.

Table 10: CART factor importance list

Factor	k = 1 or 2 5 var	k = 1 or 2 10 var	k = 1 or 2 15 var	k = 2 or 3 5 var	k = 2 or 3 10 var	k = 2 or 3 15 var
k	*27.66*	*24.99*	*22.26*	*17.02*	*16.15*	*16.49*
N_1	18.80	16.16	14.33	7.79	7.21	7.34
N_2	0.27	0.16	0.14	8.68	8.35	8.51
N_3	–	–	–	0.88	0.90	0.91
Total N (N_T)	18.78	16.09	14.33	9.09	8.65	9.01
N_2:N_1	–	–	–	3.27	3.40	3.36
N_3:N_1	–	–	–	1.17	1.13	1.23
N_3:N_2	–	–	–	2.32	2.68	2.64
N_1:N_T	1.58	1.48	1.47	3.22	3.37	3.26
N_2:N_T	–	–	–	3.11	3.30	3.18
N_3:N_T	–	–	–	2.97	3.20	3.16
$\lvert M_1 - M_2 \rvert$	**32.92**	**41.12**	**47.48**	12.38	13.62	14.77
$\lvert M_2 - M_3 \rvert$	–	–	–	7.26	8.15	7.64
$\lvert M_1 - M_3 \rvert$	–	–	–	**20.84**	**19.87**	**18.49**

Note: The calculated importance of various factors in the six CARTs. The columns indicate the number of classes being compared, along with the number of identifying variables. The rows list the factors themselves and are vaguely categorized into the following: number of groups, sample size(s), ratio between sample sizes, and the distances between group means. These scores have been scaled so that they sum to 100. The most important factor is bolded, and the second most important italicized.

The most important factor is bolded for each column. Notably, the factor that was most influential across conditions was always related to distance, followed by the true number of classes. The importance of sample size was generally smaller than these other factors, with the total sample size being the most important factor of those related to sample sizes.

CHAPTER 4

EMPIRICAL EXAMPLE

After analyzing the performance of the LI3P relative to the other techniques using a simulated dataset, the LI3P was applied to an empirical sample published by McLaughlin et al. (2020).

Empirical Background

Acknowledging a gap in the literature about the motivations for individuals who leave organized Christianity, McLaughlin et al. (2020) used a series of LPA models to determine if subgroups exist within formerly religious individuals (termed "religious dones") based on a variety of measured behaviors, beliefs, and attitudes.

Empirical Method

Empirical Participants and Procedure. To do this, researchers used the online survey platform Qualtrics to collect survey data in the United States, Hong Kong, and Netherlands (N = 3071), and 643 participants identified they were formerly religious (US n = 206; NED n = 288; HK n = 149). This sample consisted of 51.2% females (48.6% males, seven did not report) and had ages ranging from 18 to 87 (M = 44.91, SD=16.35; 31 did not report).

Empirical Measures. Religious identity was determined using a categorical three-item response scale (e.g., "*I was formerly religious, but no longer identify as religious*"). Current religious belief was assessed using a dichotomous item assessing belief in the existence of God, followed by a seven-point Likert scale assessing *commitment* to their beliefs ranging from 1 ("*Not very committed*") to 7 ("*Extremely committed*") and a 101-point Likert scale assessing *certainty* in their belief ranging from 0 ("*Not certain at all*")

to 100 ("*Extremely certain*"). Religious behaviors were assessed using an 11-point Likert scale asking how often participants engage in religious activities and other religious individuals with responses ranging from 0 ("*Never*") to 100 ("*Extremely frequently*"), in increments of 10. Finally, religious attitudes were assessed using two items regarding attitudes toward religion and religious individuals using an 11-point Likert scale with response options that range from -100 ("*Extremely negative*") to 100 ("*Extremely positive*"), in increments of 20.

Empirical Results

After fitting LPA models in M*plus* (version 8.3), McLaughlin et al. (2020) found the standard approaches yielded contradicting results: the LMR and entropy suggested two classes, whereas the BIC, the saBIC, and the BLRT suggested three classes. The specific values for each of these are reported in Table 16, along with the results of the LI3P. Notably, the models with a higher number of classes failed to converge, even after allowing for more starts, a result that was not altogether unsurprising given the fact that many over-fitting models failed to converge in our simulation.

The researchers decided to use two classes based on these results and theoretical interpretability: one class who no longer engaged in religious activities and had neutral

Table 11: LPA fit information for religious data

Class Solution	-2LL	BIC	SaBIC	Entropy	BLRT	LMR	LI3P
1 class	33312.18	33402.7	33358.25	–	–	–	–
2 classes	32529.16	32671.41	32601.56	0.91	<.0001	<.0001	0.29
3 classes	32213.28	32407.26	32312.01	0.81	<.0001	0.342	0.12
4 classes	Did not converge	Did not converge	Did not converge	0.6	Did not converge	Did not converge	Did not converge
5 classes	Did not converge	Did not converge	Did not converge	0.84	Did not converge	Did not converge	Did not converge

attitudes toward religion, and a second class that still held positive attitudes of religion and engaged in religious activities to some extent. The LI3P values indicated that using two classes instead of one was just shy of being a large effect size, as defined by Grimm et al. (2021), whereas moving from a two to a three-class model was a medium effect. We would interpret this as evidence that the three-class model constitutes a medium sized improvement in model fit compared to the two-class model.

CHAPTER 5

DISCUSSION

This study considered the performance of the LI3P measure relative to the established class enumeration techniques AIC, SaBIC, and aLMR-LRT. These methods were applied to analyze LPA models fit to data generated under a variety of conditions, including sample size, class separation, number of indicators, and the number of classes, with values generated based on past research and guidelines. After evaluating the performance in a simulation, the LI3P was then applied to an empirical dataset to demonstrate its use.

Discussion of Simulation Results

The established approaches performed largely as expected. The AIC erred on the side of more classes in the two-class case, though this was reversed in the three-class case, potentially due to the range of class distances selected. The SaBIC was highly accurate in the one-class case, and tended to choose fewer classes in the two- and three-class cases. The aLMR-LRT performed similarly, notably almost achieving perfect accuracy in the one-class case. To compare the LI3P to these measures in this instance, a cutoff score was used. Used this way, the LI3P's performance was in line with the existing measures – exceeding each in some conditions, but not in every condition. The ARI summarizes the accuracy across conditions, revealing the LI3P was more accurate than the AIC but less accurate than the SaBIC or aLMR-LRT overall.

The crosstables reveal that the LI3P tended to agree with the aLMR-LRT and SaBIC when 10 or more variables were used. In the five-variable condition, the LI3P tended to select relatively more classes. The LI3P had a peculiar inverse relationship with

the AIC in relation to the number of variables, with decreasing agreement as the number of variables increased; this may be because as the number of variables increases the AIC preferred solutions with relatively more classes. Furthermore, this reveals an interesting association between these methods: the LI3P was not consistently better or worse than any of these methods across the conditions, but its performance was based on a variety of factors. Knowing these factors may be the key to its successful use.

The scatterplots revealed numerous interesting trends in the two-class case. Depending on the number of variables present to indicate a class, there seems to be a linear relationship between sample size and class separation for the SaBIC. As is typical, for a given distance, increasing the sample size seems to be a reliable way to increase accuracy. The aLMR-LRT behaved similarly, but with slightly less accuracy. The AIC, conversely, had no discernable pattern, suggesting its accuracy was not as directly related to sample size or distance. The LI3P did not have an interaction between sample size and distance, evident by the apparent lack of a slope in the plots, but there were direct effects of both sample size and distance, contingent upon the number of variables.

Examining this more closely, the LI3P performed poorly in the five-variable condition, which was suggested by the cross tables. Using 10 or 15-variable greatly improved accuracy, at least in terms of predictability. Accuracy was almost guaranteed for data with a sample size greater than 400 and a distance greater than .5 in the 15-varaible condition. This is somewhat reassuring, meaning that to correctly determine the number of classes, sample size doesn't seem to matter after 400 samples. However, class distance also mattered, which was not as helpful because this information if not known prior to performing the analysis. By contrast, for the SaBIC and aLMR-LRT which had

an interaction effect, the user is at least reassured that by adding more individuals, the statistical power increases. The graphs seem to suggest that using the LI3P might be more useful for situations where large samples are harder to obtain, but less useful for larger samples, where other methods may achieve greater accuracy.

It is worth reiterating that these scatterplots only illustrate the two-class condition. Different trends in the three-class condition, such as the LI3P having the best accuracy in the five-variable condition and the AIC having the best accuracy overall, suggests that the trends observed in the two-class condition may not generalize. Unfortunately, these trends that are difficult to observe visually using scatterplots due to the increased dimensions of the data.

The regression analyses partially refute the claim that the LI3P is immune to the effects of sample size. Both the regression point-estimates and the squared semi-partial correlations in the three-class condition suggest that sample size can influence the LI3P values. While unfortunate, this is not entirely unexpected; using the same method, even the SaBIC, which supposedly adjusts the BIC for sample size and arguably performed the best for the simulation study, had squared semi-partial correlations of 0.29 to 0.50 for the sample sizes across the conditions. However, only the distance variables had large effect sizes for the LI3P in the three-class condition. Furthermore, in the two-class condition, only the distance manipulation had a sizeable effect.

The CART analyses largely reaffirm the regression results. For both the one-vs-two and two-vs-three class comparison, the trees reveal that the initial nodes were splits on the distance and class manipulations. Table 10 revealed that the most important factor across the conditions was a distance variable, though which distance variable changed,

followed by the number of classes. The importance of the distance variables, particularly the distance between the first and third classes, is highlighting the importance of class separation in distinguishing the classes. Simply put, the further apart the classes, the easier it is to determine the correct number of classes.

Discussion of Religious Data Results

The empirical example demonstrates a case where the two most accurate methods according to this study, the LMR and SaBIC, were at odds with each other. The LI3P indicated a medium improvement in fit. This would help directly address the concerns of the authors that the class enumeration technique scores did not change as greatly when comparing the one-to-two-class model to the two-to-three-class model, suggesting that the change in model fit is not so small as to be negligible. This, in conjunction with the other class enumeration methods, could be used to justify using a three-class solution.

However, it is worth acknowledging the logic the authors make that this was an exploratory analysis and the interpretability was clearer for two classes. They admittedly want to be more conservative in their approach of determining the number of latent classes in their study. Hopefully, continued research in this field will help clarify these results, perhaps by increasing the sample size and the number of measures used.

Future Directions and Concluding Remarks

Future directions might seek to evaluate the effectiveness of the LI3P using an adjustable cutoff value. For example, in our study, this study used a cutoff criterion of 0.014 because it was the value at the 95^{th} percentile for the 15-variable, one-class condition. However, as demonstrated by the study findings, the number of indicators affect the LI3P performance, so employing an approach that considers the number of

predictors could be useful. In our data, this would be 0.022 for the 10-indicator case, and 0.044 for the five-indicator case. A different approach but with a similar goal would be to use scree plots. Typically used in principal and confirmatory factor analysis, this technique could be applied here to simulate data under the null, then compare the observed LI3P value to the values produced by the null.

Additionally, further research may explore LI3P performance in LCAs and GMMs more thoroughly, as these are related and popular techniques not explored here. The class enumeration techniques discussed here are used in these models, though their performance is not always the same, so exploring the application of the LI3P in those models would be recommended over generalizing these results to them.

Overall, these results suggest that while the LI3P is a potentially useful tool in a researcher's statistical toolkit, even though it did not perform as well as originally expected. While sample size wasn't a primary factor in its calculation, it does play a small role in certain conditions, meaning the LI3P isn't entirely "free" of the influence of sample size. That being so, the regression analyses reveal that it is more reliant on the desired factors of class separation and the true simulated number of classes. The LI3P performed better than the AIC in terms of overall accuracy, but not as well as the SaBIC or LMR. The LI3P tended to favor fewer classes when 10 or more variables were used, which was not uncommon; all of the class enumeration techniques tended to favor fewer classes than the correct solution, likely due to the challenging sample size and class distance conditions used. The LI3P may be best used when the sample is relatively small, and use 10 indicators to balance accuracy. When used with the existing methods, the LI3P can help researchers make informed decisions by providing context by reporting

scores that reflect distance and the number of classes.

REFERENCES

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *2nd International symposium on information theory*. Editors B. N. Petrov, and F. Csáki (Budapest, Hungary: Akadémiai Kiadó), 267–281.

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, *40*(2), 235–243. https://doi.org/10.1509/jmkr.40.2.235.19225

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. https://doi.org/10.1007/BF02294361

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 45-70.

Finch, H. (2015). A comparison of statistics for assessing model invariance in latent class analysis. *Open Journal of Statistics*, *05*(03), 191–210. https://doi.org/10.4236/ojs.2015.53022

Gonzalo, J., & Pitarakis, J. Y. (2002). Lag length estimation in large dimensional systems. *Journal of Time Series Analysis*, *23*(4), 401–423. https://doi.org/10.1111/1467-9892.00270

Grimm, K. J., Houpt, R., & Rodgers, D. (2021). Model fit and comparison in finite mixture models: A review and a novel approach. *Frontiers in Education, 6*, 613645. https://doi.org/10.3389/feduc.2021.613645

Hubert, L., & Arabie, P. (1985). Comparing partitions, *Journal of Classification*, *2*(1), 193-218.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Jeffries, N. O. (2003). A note on "Testing the number of components in a normal mixture." *Biometrika*, *90*(4), 991–994. https://doi.org/10.1093/biomet/90.4.991

Liu, M., & Lin, T. I. (2014). A skew-normal mixture regression model. *Educational and Psychological Measurement*, *74*(1), 139–162. https://doi.org/10.1177/0013164413498603

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Magidson, J., & Vermunt, J. K. (2002). A nontechnical introduction to latent class models. *DMA Research Council Journal*. Available from: http://www.statisticalinnovations.com/articles/lcmodels2.pdf.

Maynard, B. R., Salas-Wright, C. P., Vaughn, M. G., & Peters, K. E. (2012). Who are truant youth? Examining distinctive profiles of truant youth using latent profile analysis. *Journal of Youth and Adolescence*, *41*(12), 1671–1684. https://doi.org/10.1007/s10964-012-9788-1

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*(1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115

McLachlan, G., and Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons, 419.

McLaughlin, A. T., Van Tongeren, D. R., Teahan, K., Davis, D. E., Rice, K. G., & DeWall, C. N. (2020). Who are the religious "dones?": A cross-cultural latent profile analysis of formerly religious individuals. *Psychology of Religion and Spirituality*. https://doi.org/10.1037/rel0000376

Morgan, G. B., Hodge, K. J., & Baggett, A. R. (2016). Latent profile analysis with nonnormal mixtures: A Monte Carlo examination of model selection using fit indices. *Computational Statistics & Data Analysis*, *93*, 146–161. https://doi.org/10.1016/j.csda.2015.02.019

Muthén, L. K., & Muthén, B. O. (1998–2014). *M*plus *user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*(2), 463–469. https://doi.org/10.1111/j.0006-341X.1999.00463.x

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Olivera-Aguilar, M., & Rikoon, S. H. (2018). Assessing measurement invariance in multiple-group latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 439–452. https://doi.org/10.1080/10705511.2017.1408015

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: A Monte Carlo simulation. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 616–639. https://doi.org/10.1080/10705511.2013.824780

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343. https://doi.org/10.1007/BF02294360

Swanson, S. A., Lindenberg, K., Bauer, S., & Crosby, R. D. (2012). A Monte Carlo investigation of factors influencing latent class analysis: An application to eating disorder research. *International Journal of Eating Disorders*, *45*(5), 677–684. https://doi.org/10.1002/eat.20958

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 640–657. https://doi.org/10.1080/10705511.2013.824781

Therneau, T., & Atkinson, B. (2019). rpart: Recursive partitioning and regression trees (4.1-15) [Computer software manual]. https://CRAN.Rproject.org/package=rpart.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, NC: Information Age Publishing.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*(2), 307–333. https://doi.org/10.2307/1912557

Wang, Y., Kim, E., & Yi, Z. (2021). Robustness of latent profile analysis to measurement noninvariance between profiles. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164421997896

Woodroofe, M. (1982). On model selection and the arc sine laws. *The Annals of Statistics*, *10*(4). https://doi.org/10.1214/aos/1176345983

Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, *50*(4), 1090–1104. https://doi.org/10.1016/j.csda.2004.11.004

Yang, C. C., & Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification*, *24*(2), 183–203. https://doi.org/10.1007/s00357-007-0010-1

Zhang, J., Zhang, M., Zhang, W., & Jiao, C. (2014). Model selection for complex multilevel latent class model. *Communications in Statistics - Simulation and Computation*, *43*(4), 838–850. https://doi.org/10.1080/03610918.2012.718836